DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving

Year: 2024 | *Citations: 378* | *Authors: Yinmin Zhong, Shengyu Liu, Junda Chen*

Abstract

DistServe improves the performance of large language models (LLMs) serving by disaggregating the prefill and decoding computation. Existing LLM serving systems colocate the two phases and batch the computation of prefill and decoding across all users and requests. We find that this strategy not only leads to strong prefill-decoding interferences but also couples the resource allocation and parallelism plans for both phases. LLM applications often emphasize individual latency for each phase: time to first token (TTFT) for the prefill phase and time per output token (TPOT) of each request for the decoding phase. In the presence of stringent latency requirements, existing systems have to prioritize one latency over the other, or over-provision compute resources to meet both. DistServe assigns prefill and decoding computation to different GPUs, hence eliminating prefill-decoding interferences. Given the application's TTFT and TPOT requirements, DistServe co-optimizes the resource allocation and parallelism strategy tailored for each phase. DistServe also places the two phases according to the serving cluster's bandwidth to minimize the communication caused by disaggregation. As a result, DistServe significantly improves LLM serving performance in terms of the maximum rate that can be served within both TTFT and TPOT constraints on each GPU. Our evaluations show that on various popular LLMs, applications, and latency requirements, DistServe can serve 7.4x more requests or 12.6x tighter SLO, compared to state-of-the-art systems, while staying within latency constraints for>90% of requests.